The Case-Control Study as Data Missing by Design: Estimating Risk Differences

Sholom Wacholder

There are advantages to viewing the case-control design as a missing-data problem instead of as a sampling problem. In the simplest setup, cases are those members of a population who develop disease; controls can be a small random sample of the large number who do not; and covariates, including exposures and other important variables, are available only for cases and controls and are assumed to be missing at random for the remaining large fraction of the population. This approach allows estimation of the joint distribution of all variables in the population. Thus, when the size of the population is known, analysis is not restricted to logistic and other multiplicative intercept models. Methods based on this approach can obtain estimates and confidence intervals for parameters representing the effect of exposure on disease, with multivariate adjustment for other factors. Thus, case-control data can be used to estimate the risk difference, a parameter with great public health value. The missing-data perspective offers an additional advantage by linking the "study base principle" of control selection with the statistical concept of "missing at random." As an illustration, I use a subset of data from a case-control study to obtain estimates of the difference between annual risk of bladder cancer for various levels of smoking and lifetime nonsmokers, adjusted for occupational exposure. (Epidemiology 1996;7:144–150)

Keywords: bias, biometry, epidemiologic methods, logistic regression, missing data, study design.

In essence, the analysis of a case-control study should translate differences in exposure between cases and controls into estimates of parameters that compare risk or rate of disease. This translation requires estimation of parameters of a *prospective* model of risk as a function of exposure, despite sampling that is conditional on disease.[1] Cornfield[2] showed that the odds ratio, which can be estimated from case-control data, could be used to approximate the risk ratio, with its prospective interpretation. The main line of thinking about the analysis of case-control data has extended this idea.[1,3–6]

The relative risk or the rate ratio has remained the parameter of interest in case-control studies, despite long-standing appreciation that exposure-specific rates of disease can be estimated if the fractions of incident cases and of noncases in the cohort selected as controls are known.[2,5,7–13] Several recent papers have incorporated the sampling from the cohort in developing the analysis of case-control studies.[14–20]

If *risks* and *risk differences* can be estimated from many case-control studies, why aren't they? In part, statistical theory and software are more developed for estimating odds ratios and their standard errors. In part, the paradigm of translating exposure odds ratios into disease odds ratios still holds sway. The paradigm of the case-control study as a cohort study with data missing at random may encourage more flexible and more informative analysis.

The Case-Control Study as a Missing-Data Problem

A case-control study can be seen simply as a study in a cohort with some missing exposure data. The investigator studies controls, rather than everyone in the underlying cohort or study base, to learn about the distribution of exposure in the whole cohort.[21,22] Indeed, the case-control data can be used to reconstruct the data from the whole cohort with but one additional piece of information—the crude disease rate, or, equivalently, when all cases are identified, the total numbers of subjects at risk. That is, case-control data can be recast as a missing-data problem. Typically, the crude rate in the population or the count of the denominator will be available when all cases have been identified from a well-defined population during a fixed period of time, as from a population-based cancer registry such as the Surveillance, Epidemiology and End Results (SEER) program.

This missing-data problem exhibits two unusual characteristics: the fraction missing can be enormous, and the "missingness" is intentional. By contrast, in most study situations, investigators assiduously avoid or minimize missingness.

Statistical methods for handling missing data depend on the mechanism that generated the missingness. Analysis restricted to units with complete data is generally valid only under the strong assumption called *missing*

From the Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, EPN 403, 6130 Executive Boulevard, Bethesda, MD 20892-7368.

Submitted April 28, 1995; final version accepted September 15, 1995.

completely at random (MCAR),[23] that the (sometimes unknown) value of a variable for an individual is independent of whether or not it is observed. Most of the statistical work has been done under the milder assumption called *missing at random* (MAR),[23] where the required independence of the value of the variable and whether it is observed is relaxed to be conditional on the values of a variable that is known for everyone; here, the conditioning is on disease status. Handling missing data that is generated by a mechanism that cannot be assumed to be missing at random and is not well understood remains problematic, as discussed by Vach and Blettner[24] for missing confounders.

The epidemiologist's assumption of appropriate selection of cases and controls from the cohort or base is equivalent to an assumption of missing at random. That is, the epidemiologist assumes that values of covariates for those who develop disease do not depend on whether the person is included as a case in the study, and the values of those who do not develop disease are assumed to be independent of whether or not the person is included as a control. Logistic regression for case-control data with appropriately selected cases and controls yields consistent estimates of the odds ratios,[5,18] even though the missingness is *at random*. This is a remarkable exception to the general rule requiring missingness that is *completely* at random for valid analysis using only complete units; it applies only where the *link* (functional form of the risk modeled by the covariates) is the odds, or a transformation of the odds, such as the logit (logarithm of the odds).[18,25] The approach developed here is more general; it treats a case-control study as a missing-at-random problem and allows estimation of additional risk parameters, such as absolute risk or risk difference.

The missing-data approach also offers conceptual and practical benefits. It emphasizes the relation of a case-control study with its underlying cohort and with the hypothetical study that could have been done using the entire cohort and exactly the same cases. It links the epidemiologic literature about case-control studies with the formal idea of data missing at random. The approach makes clear the relation between the *study base principle*,[21,22] one of the three underlying principles behind control selection,[22] and *missing at random*,[23] thereby providing a theoretical framework to common epidemiologic practices. Finally, it offers a powerful tool to address questions about the analysis of case-control studies, particularly those with complex sampling.[18,20]

Estimating the Risk Difference from a Case-Control Study

Standard missing-data theory can be used to show that a parameter that can be estimated from a cohort study also can be estimated from a case-control study under the missing-at-random assumption. This approach has been used by Benichou and Wacholder[12] to develop estimators of the absolute risk of disease under a logistic model when the study identifies all incident disease during a fixed time interval in a population of known size. I show below how the probability of disease itself, or transformations of the probability in addition to the logit, can be modeled and estimated as a function of risk factors collected in the case-control study.

What practical advantages does the risk-difference parameter offer? First, it is the natural parameter for public health interpretation.[26] Second, it encourages a realistic comparison of effects in strata with large differences in baseline risk, as in a study of differences in the effect of alcohol on risk of cancer of the esophagus between blacks and whites. Third, it allows the total effect of a single risk factor to be evaluated across endpoints with different levels of risk. For example, evaluation of the benefits and risks of hormone replacement therapy for postmenopausal women requires consideration of possible effects on risk of heart disease, breast and endometrial cancer, and osteoporosis. Comparing these effects using risk differences is much simpler than on a relative scale. Fourth, the effects of a continuous risk factor or the joint effects of several risk factors might be better described by a risk-difference model. In this paper, I outline how in some circumstances one can use case-control data to estimate risk differences in a multivariate setting and how to obtain confidence intervals.

Theoretical Results

Weinberg and Wacholder[18] use related ideas to justify prospective analysis of case-control studies with discrete covariates. A fully general theoretical result that incorporates continuous covariates, as do Prentice and Pyke,[5] might be obtained using the approach of Wild.[17] The results of Weinberg and Wacholder[18] are themselves generalizations of those of Anderson[4] to include *multiplicative intercept models*,[25] that is, models with links that are functions of the odds but do not extend to parameters that are not dependent on the odds, such as the risk difference. When the outcome is rare, the odds and the probability itself are close, so results from modeling of the odds are approximately equal to those from modeling the probability.

The Missing-Data Paradigm for Case-Control Studies

Every case-control study can be considered as arising from a base or a cohort.[22] In the simplest situation, cases from a "cumulative incidence"[27] study consist of everyone in a cohort who develops disease during a fixed period of time, and controls are a randomly selected subset of the cohort, excluding the cases. Covariates are collected from cases and controls; the data from the other members of the cohort are *missing by design*. If the controls have the same distribution of exposure as the noncases in the population from which cases are drawn, the study base criterion[22] is met, or the data are missing at random. Thus, standard missing value methods can be applied.[23] If the study base criterion is not met, the missingness is not at random, and use of either standard or missing-data analysis for a case-control study would generally result in biased estimation of risk parameters.[22]

The Likelihood Function

Assume that in the population of size N, there are N_i individuals with $D = 1$, where diagnosis of disease during the follow-up period is indicated by $i = 1$, and $i = 0$, otherwise. Then n_1 cases and n_0 controls are sampled without replacement from among the N_1 and N_0 diseased and nondiseased subjects; covariate information is obtained only from these $\Sigma_i n_i$ subjects. Denote the numbers of cases and controls with an observed discrete covariate vector value of x as $n_1(x)$ and $n_0(x)$, respectively. The full likelihood L for the entire cohort, including those for whom disease status but not covariates are known, can be factored via Bayes' rule, $\Pr(x|D) = \Pr(D|x)\Pr(x)/\Pr(D)$, to show its dependence on the risk model[17,18]:

$$
\begin{aligned}
L &= \{\Pi_i \Pr(D = i)^{N_i}\} \{\Pi_{i,x} \Pr(x|D = i)^{n_i(x)}\} \\
&= \{\Pi_i \Pr(D = i)^{N_i - n_i}\} \{\Pi_x \Pr(x)^{\Sigma n_i(x)}\} \qquad (1) \\
&\qquad \{\Pi_{i,x} \Pr(D = i|(x)^{n_i(x)}\}.
\end{aligned}
$$

The extensions required for continuous variables are developed by Wild.[17]

Case-Control Studies with a Primary Base

In a study with a *primary base*, or a "population-based case-control study," the investigator identifies all of the cases developing during a specified time interval in a fixed population. In fact, under the missing-at-random assumption, the expected value of the sufficient statistic for parameters of risk in the cohort (the *sufficient statistic* contains all of the information in the data about the parameter of interest; here, it is the table classifying all members of the cohort jointly by disease and exposure status[20]) can be obtained; therefore, any parameter that could be estimated from the full cohort can also be estimated from a population-based case-control study. Thus, missing-data methods can be used to estimate not only absolute risk but parameters such as the risk ratio or risk difference[28,29] from a model based on a nonlogistic link[30] through either a pseudo-likelihood or full-likelihood[20] approach when the variables are discrete.

Often, the information required to exploit missing-at-random methods is readily available. For instance, age-, sex-, and race-specific counts of the numbers of subjects at risk, that is, the denominators of vital statistics rates, can be obtained from census data. In fact, this approach can be seen as an extension of vital statistics modeling, incorporating demographic covariates available for everyone, together with exposure covariates only collected for cases and controls.[14-18] The method may also be useful even when the denominator counts are only approximate.

In a study with a secondary base,[21,22] such as a hospital- or registry-based study,[31] the base is implicitly defined in reference to the mechanism used to collect cases. But although hospital controls may be adequate for estimating relative risk, estimation using the techniques described here requires enumeration of the cohort, which will usually not be possible.

Estimation

Risk models with various links can be fit. *Pseudo-likelihood* estimates can be obtained by distributing the cases and controls with missing covariates into the exposure cells proportionally to the empirical distributions of observed cases and controls, respectively, and proceeding as if the resulting table were the full cohort. The standard errors reported by a packaged program are too small, however; appropriate standard errors can be obtained from the "sandwich" variance estimator described for method 2 in section ¶3.3 of Benichou and Wacholder.[12] Alternatively, the *EM* (*Expectation Maximization*) algorithm[32] could be used to obtain maximum likelihood estimates. As in the logistic case,[20] the *E* step assigns subjects with missing covariates into cells in proportion to the fitted values (in contrast to the observed values in the pseudo-likelihood). The difference is that the M step can fit the completed data using a link other than the logit.[28] The *Newton-Raphson* algorithm[33] generates maximum likelihood estimates by directly maximizing the likelihood L from Eq 1. Standard errors can be obtained by inverting the observed information matrix based on the second derivative of the likelihood L.

Estimation is simpler when the link is a function of the odds, and the numbers of diseased and nondiseased persons in the cohort are known. For calculation of the odds-difference model, I defined a generalized linear model with a link that incorporates the ratio of sampling fractions (proportions in the cohort included in the analysis) for cases to controls in the program Gauss (Aptech Systems, Kent, WA). The fitted probabilities and the derivative of the link function depend on R, the stratum-specific ratio of sampling fractions in cases and controls, that is, the quotient of the fractions of incident disease and nondiseased individuals in that stratum of the cohort whose exposure information is used in the analysis. The fitted probabilities $\hat{p}$ are calculated from the *linear predictor* ℓ (the vector of the sum of the product of the regression variables and the estimated regression coefficients) as $\hat{p} = \ell/(\ell + 1/R)$. The derivative of the link is $1/[R(1 - \hat{p})]$. Weinberg and Sandler[19,pp426-427,432] used a slightly different approach to fit an odds-difference model in GLIM (Numerical Algorithms Group, Oxford, England); I specify R, in contrast to Weinberg and Sandler,[19] who were interested in the joint effects of two factors rather than estimates of their individual effects.

Simple Hypothetical Example

A case-control study targets a specified race-sex-age stratum in a study population specified both geographically and temporally. The stratum consists of 100,000 persons; all of the 20 cases that are incident during follow-up are identified (Table 1). Exposure information is obtained from the cases along with 20 controls drawn at random from the stratum, and 10 cases and 5 controls are found

TABLE 1. Hypothetical Example for Estimating Risk Difference, Accounting for All 100,000 Persons in the Stratum

	In Stratum			Exposed			Unexposed		
	Total	Observed	Missing by Design	Total	Observed	Inferred	Total	Observed	Inferred
Incident disease	20	20	0	10	10		10	10	
No incident disease	99,980	20	99,960	24,995	5	24,990	74,985	15	74,970
Total	100,000	40	99,960	25,005	15	24,990	74,995	25	74,970
Risk	20.0			40.0			13.3		

to be exposed. There remain 99,960 (100,000 total − 20 cases − 20 controls) persons in the stratum for whom exposure information is unknown. If the controls truly are randomly drawn, then the exposure levels of these 99,960 persons are missing at random and can be expected to include 24,990 exposed and 74,970 unexposed. These estimates can now yield estimates of the counts of nondiseased subjects *in the population* at each level of exposure in the stratum and, hence, estimate the difference in risk as [10/(24,990 + 5 + 10)] − [10/(74,970 + 15 + 10)] = (10/25,005) − (10/74,995) = 26.7 per 100,000 persons per year.

When exposure is not obtained from all incident cases, its distribution can be similarly extrapolated from a set of cases who constitute a random sample of diseased subjects. Generalization to cells formed by cross-classification of several discrete covariates is straightforward.

In this example, the pseudo-likelihood approach was used for simplicity. In a more complex situation, when the risk model is not saturated in parameters, the likelihood approach, which uses expected numbers of cases and noncases based on the fitted model, rather than the observed numbers, will yield different estimates.[20]

Data Example

Methods

Analyses of data extracted from the National Bladder Cancer Study[34,35] (NBCS) are presented as examples. Readers interested in the substantive results should consult the published report.[35] In this study, all residents of SEER catchment areas diagnosed with bladder cancer in 1978 were eligible to be cases. Random digit dialing and Health Care Financing Administration records were used for selecting controls below 65 years of age and 65 years or older, respectively. In addition, population counts by age, race, and sex, used by SEER as denominators for incidence rates, were available. Cases and controls who agreed to participate provided information on several variables considered to be possible risk factors for bladder cancer. For the analyses here, I consider only a single stratum consisting of white men age 65–79 years in the state of New Jersey. These analyses use data presented in Table 2 to examine the effects of smoking at four levels and of work in any occupation from a list of those possibly related to bladder cancer.[35]

The risk-difference model uses the identity link so that the probability itself, rather than a transformation of the probability, is fit; thus, the model parameters for covariates can represent differences in absolute risk[28] during a specified time interval. This example uses models of difference in annual risk of the form:

$$\Pr(D|S = i, E = j) = \alpha + \beta_i + \gamma_j, \qquad (2)$$

where $S = i$, $i = 0,1,2,3$, means never-smoker, former smoker, current light smoker, and current heavy smoker, respectively, and $E = j$, $j = 0,1$, indicates occupational exposure. In these models, $\beta_0 \equiv \gamma_0 \equiv 0$. In a model of the effects of smoking only, the parameter α is the risk at the baseline (never-smokers), the β_i, $i = 1,2,3$, are the differences in annual risk between level i and level 0 of smoking, and $\gamma_1 \equiv 0$. In a model also including the effect of occupational exposure, α is again the baseline risk, this time referring to nonsmokers who are not occupationally exposed, and the β_i and γ_j represent differences in annual risk at $S = i$, $i = 1,2,3$, and $E = 1$ compared with $S = 0$ and $E = 0$, respectively, with the level of the other variable held fixed. Alternatives to Model 2 would replace the probability on the left-hand side of Model 2 by an odds $[p/(1 - p)]$, logarithmic, or logit $(\log[p/(1 - p)])$ transformation. Then, the βs and γ would represent differences in the odds or in the logarithm of the risk or odds (the familiar "log-odds ratio"), respectively.

Results

Tables 3–5 display estimates of parameters from the risk-difference, odds-difference, and odds ratio models. Because the univariate model is saturated, identical es-

TABLE 2. Exposure and Smoking Status of White Men Age 65–79 Years in the State of New Jersey, from National Bladder Cancer Study Data[34,35]

		Smoking Status by Occupational Exposure							
		Yes				No			
	Unknown Exposures	Never	Former	Light	Heavy	Never	Former	Light	Heavy
Diseased	195	24	46	5	30	49	95	16	71
Nondiseased	90,441	94	125	14	41	90	152	20	71

TABLE 3. Maximum Likelihood Estimates (× 10^5) of Baseline Risks and of Risk Differences for Annual Risk of Bladder Cancer for White Men Age 65–79 Years in New Jersey, from National Bladder Cancer Study Data[34,35]

	α*	β_1*	β_2*	β_3*	γ_1*
Smoking only					
Estimates	416.	117.	230.	525.	
Standard errors	53.	79.	187.	138.	
Smoking and occupation					
Estimates	271.	106.	175.	487.	292.
Standard errors	55.	74.	177.	135.	72.

* As defined in Model 2 in the text.

timates are obtained from full- and pseudo-likelihood,[12] the fitted values will be the same from any link, and the parameter estimates and standard errors from a fit using one link can be obtained from the estimates of another. The pseudo-likelihood estimates (not shown) for the two-variable model are similar to those from the full-likelihood, as noted by Benichou and Wacholder[12] for a larger subset of the bladder cancer study. The estimates from the odds-difference model are slightly larger than from the risk-difference model, since an odds always exceeds the corresponding probability.

Results from fitting the risk-difference (Model 2) indicate that the annual risk of bladder cancer in heavy smokers in this stratum is increased by almost 500 per 100,000 compared with lifetime nonsmokers, whereas work in one of the included occupations increases the annual risk of bladder cancer by nearly 300 per 100,000. Thus, the annual risk for an occupationally exposed heavy smoker is estimated to be about 800 per 100,000 higher than for an unexposed man who never smoked. The odds ratio estimates from Model 3 are 2.2 for heavy smoking and 1.7 for occupational exposure, respectively. By adding occupation to a model with only smoking, the deviance changes by 14.8 and 15.8 in the risk-difference and log-odds models, respectively. The baseline models have the same deviance, so the difference in deviances between the two models is only 1.0, and the data do not provide a firm basis for choosing between them. Even though the deviances are similar, some of the estimated differences in exposure-specific risk (Table 6) can be substantial: up to 15% for heavy smokers who are not occupationally exposed.

TABLE 4. Odds-Difference Estimates (× 10^5) of Baseline Risks and of Risk Differences for Annual Risk of Bladder Cancer for White Men Age 65–79 Years in New Jersey, from National Bladder Cancer Study Data[34,35]

	α*	β_1*	β_2*	β_3*	γ_1*
Smoking only					
Estimates	418.	118.	233.	532.	
Standard errors	55.	76.	180.	135.	
Smoking and occupation					
Estimates	272.	107.	177.	493.	295.
Standard errors	57.	75.	176.	139.	75.

* As defined in the text.

TABLE 5. Estimates of Parameters of Model with Logit Link for Annual Risk of Bladder Cancer for White Men Age 65–79 Years in New Jersey, from National Bladder Cancer Study Data[34,35]

Smoking and Occupation	α*	β_1*	β_2*	β_3*	γ_1*
Estimates	−5.80	0.23	0.39	0.77	0.55
Standard errors†	0.16	0.17	0.31	0.20	0.14
Odds ratios		1.3	1.5	2.2	1.7

* As defined in the text.
† Of the log-odds ratio.

Discussion

I have shown that a case-control study always can be viewed, and sometimes can be analyzed, as a cohort study with missing data. When a case-control study is appropriately designed and implemented, it parallels a hypothetical cohort study performed in the same setting, because the data not collected in the case-control study are missing by design and missing at random. When the study is poorly designed or implemented, the data *are not* missing at random, and the validity of standard analyses, as well as those explicitly based on the missing-at-random assumptions, would be questionable.

Efficiency of Analysis

Even though the analysis illustrated here incorporates the crude rate of disease, it does not provide more efficient inference on the odds ratio,[18] which depends on proportions rather than absolute numbers. Thus, the odds ratio estimates for smoking and occupation and their standard errors shown in Table 4 are identical to those obtained from ordinary logistic regression analysis, as the reader can verify.

Control and Case Selection

The missing-data viewpoint provides an attractive formal statistical framework for discussing practical problems in control selection. The two most common problems,[22] failure to obtain a random sample from the study

TABLE 6. Fitted Probabilities (× 10^5) from Models of Annual Risk of Bladder Cancer for White Men Age 65–79 Years in New Jersey, from National Bladder Cancer Study Data[34,35]

Covariate Level		Model		
			Smoking and Occupation	
Smoking	Occupational Exposure	Smoking	Risk Difference	Logistic
Never	No	416.	271.	304.
Never	Yes	416.	563.	577.
Former	No	533.	378.	381.
Former	Yes	533.	669.	660.
Light	No	647.	446.	448.
Light	Yes	647.	738.	774.
Heavy	No	941.	758.	652.
Heavy	Yes	941.	1049.	1126.

base, particularly when there is no roster available, and refusal to participate, can make the missing-at-random assumption tenuous, even for case-control studies that are claimed to be population based.

Case ascertainment can be as troubling as control selection. Nonmultiplicative risk models (links other than logistic and logarithmic for odds ratio or risk ratio parameters, respectively[28]) require complete case identification; for logistic risk models, cases missing at random will have a slight effect on estimates of the effects of covariates to the extent that missed cases contaminate the count of the nondiseased. But under an assumption of missingness at random, unidentified cases result in a proportional downward bias in estimates of rates of disease, even for multiplicative models. Even if missingness of cases is unrelated to covariates, estimates of risk difference will also be proportionally reduced. Note that, at least for cancer, the cases available to the study are often identified through a registry such as SEER and would be the basis of the numerator for calculation of reported incidence rates. Thus, any inaccuracies in the population counts will also manifest themselves as errors in vital statistics rates and as bias in estimates of absolute, but not relative, risk; in addition, there will be bias in estimates of parameters of exposure effects when using other links. Thus, if there was underidentification of cases unrelated to exposure, the estimates of annual risk in Table 6 will be biased downward by a fixed fraction.

Identified cases for whom covariates are not collected, possibly because of death before interview or exclusion due to lack of histologic confirmation, can lead to violation of the missing-at-random assumption if their exposure distribution is different from that of other cases. For example, patients seen at rural primary care hospitals may be less likely to have a work-up including histology than patients at an urban tertiary care hospital. The missing-at-random assumption may then be violated in an occupational study.[31]

Risk-Difference and Other Nonlogistic Models

This is, I believe, the first time that the parameters of a multivariate risk difference or, in fact, any multivariate model outside of the class of multiplicative intercept models[18,25] have been obtained from a case-control study. The missing-data approach offers a major advantage by allowing estimation of absolute risk and risk-difference parameters from a case-control study when the total number of subjects at risk is available. These parameters measure units of risk within a time period and, therefore, will often measure impact on public health far better than unitless ratio parameters. Use of risk-difference estimates would illuminate rather than obscure the difference in importance between control of an exposure with a relative risk of 3 for a rare cancer and one with a relative risk of 3 for coronary heart disease. Furthermore, risk models with other links can be explored; for example, Weinberg[36] has suggested using the og-complement link to obtain a *health ratio* when assessing independence among causal factors.

The missing-at-random approach provides new opportunities for examining the joint effects of more than one risk factor in case-control data. When multiple strata of the bladder cancer dataset are considered, with vastly different baseline incidence rates (data not shown), the logistic model fits much better than the risk difference, in keeping with the observation of Breslow and Day that "the epidemiology of cancer . . . provides empirical reasons for choosing relative risk as the natural measure of cancer and exposure."[6,p68] Nonetheless, other measures of association may be appropriate for other diseases, and there may be situations in which an alternative form of model of joint effects is appropriate, even for some sites of cancer. Although one could employ software such as Egret (Statistics and Epidemiology Research Corporation, Seattle, WA) or Epicure (Hirosoft International Corporation, Seattle, WA) to compare multiplicative and additive relative risk models of the joint effects of two exposures, the missing-at-random approach offers a much broader class of options.

When the outcome is rare, use of the odds-difference as an approximation for the risk difference offers the simplest approach. Software for fitting generalized linear models, as in GLIM (Numerical Algorithms Group, Oxford, England) or SAS (SAS Institute, Cary, NC), is required. Use of an odds, identity, or other nonlogistic link with binomial data, however, can cause computing and inference problems when the fitted value of each probability is not guaranteed to lie between 0 and 1.[19,20]

In multiplicative intercept models, a computational simplification arises since the sampling fractions for cases and controls can act as expansion estimators applied, respectively, to the numerator p and the denominator $1 - p$ of the odds. The computation of maximum likelihood estimates for models outside the multiplicative intercept class is nontrivial and requires special software even with discrete covariates. There are usually many nuisance parameters. Applications with more than one stratum and with continuous covariates will involve challenging computing problems even as the theoretical problems are solved. Use of pseudo-likelihood estimates, which are computable using software for generalized linear models,[19,28] will avoid the problem of nuisance parameters but not of restricted fitted probabilities. Nevertheless, calculations beyond standard computer output are required to obtain confidence intervals when using pseudo-likelihood.

Other Applications of the Missing-Data Approach in Case-Control Studies

The concept of data missing by design has been extended to more complex case-control sampling designs.[20,37] In frequency matching, two-stage designs,[15,19,20,38-40] and the partial questionnaire design,[37] some variables are not collected from a subset of cases and controls, determined randomly and perhaps depending on the level of a variable that has been collected on everyone. Again, the data are missing by design, and the missing-at-random assumption will hold in theory. Here, a roster of subjects in the sampling frame is available, so the missing-at-random assumption is likely to hold in practice as well. Methods of analysis for two-stage designs can accommodate matching on a variable collected only from potential cases and controls,[20,41] as when choosing subjects for assessment of residential radon exposure on the basis of smoking information; matching on a variable that is known for everyone in the study base (typically, variables such as age, race, sex, and region) can be handled by forming strata based on these variables and applying the methods developed in this paper.

Nested case-control and case-cohort studies can also be thought of as full cohort studies with missing data. The likelihoods used for cohort, case-control, and case-cohort studies are identical except for the composition of the comparison sets for the cases[42]; the comparison sets for each design constitute a random sample of the risk set in the corresponding cohort study.[42] This parallelism suggests that decisions about exclusion of subjects and calculation of time-dependent covariates for nested case-control and case-cohort studies should be made exactly as for the cohort study.

Validity of Case-Control Study Design

Finally, the missing-data framework makes explicit the relation between a case-control study and its underlying cohort or study base. This structure demonstrates simply, even elegantly, the theoretical validity of the case-control design. Furthermore, it reveals the conditions critical to translating validity in theory into validity in practice. The missing-at-random perspective emphasizes that the design, implementation, and criticism of case-control studies ought to focus on case ascertainment, control selection, and data quality rather than on a misguided characterization of the case-control study as intrinsically unreliable.

Acknowledgments

I thank Patricia Hartge and Clarice Weinberg for useful comments both about the ideas underlying this paper and about how to express them.

References

1. Breslow NE, Powers W. Are there two logistic regressions for retrospective studies? Biometrics 1978;34:100–105.
2. Cornfield J. A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast and cervix. J Natl Cancer Inst 1951;11:1269–75.
3. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–748.
4. Anderson JA. Separate sample logistic discrimination. Biometrika 1972;59: 19–35.
5. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979;66:403–411.
6. Breslow NE, Day NE. Statistical Methods in Cancer Research. vol. 1. The Analysis of Case-Control Studies. IARC Scientific Pub. No. 32. Lyon: International Agency for Research on Cancer, 1980.
7. Doll R, Hill AB. Smoking and carcinoma of the lung: preliminary report. BMJ 1950;2:739–748.
8. MacMahon B. Prenatal x-ray exposure and childhood cancer. J Natl Cancer Inst 1962;28:1173–1191.
9. Cole P, Monson RR, Haning H, Friedell GH. Smoking and cancer of the lower urinary tract. N Engl J Med 1971;284:129–134.
10. Greenland S. Multivariate estimation of exposure-specific incidence from case-control studies. J Chron Dis 1981;34:445–453.
11. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989; 81:1879–1886.
12. Benichou J, Wacholder S. A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. Stat Med 1994;13:651–661.
13. Benichou J, Gail MH. Methods of inference for estimates of absolute risk from population-based case-control studies. Biometrics 1995;51:182–194.
14. Fears TR, Brown CC. Logistic regression methods for retrospective case-control studies using complex sampling procedures. Biometrics 1986;42:955–960.
15. Breslow NE, Zhao LP. Logistic regression for stratified case-control studies. Biometrics 1988;44:891–899.
16. Scott AJ, Wild CJ. Fitting logistic regression models in stratified case-control studies. Biometrics 1991;47:497–510.
17. Wild CJ. Fitting prospective regression models to case-control data. Biometrika 1991;78:705–717.
18. Weinberg CR, Wacholder S. Prospective analysis of case-control data under general multiplicative-intercept risk models with biased sampling. Biometrika 1993;80:461–465.
19. Weinberg CR, Sandler D. Randomized recruitment in case-control studies. Am J Epidemiol 1991;134:421–432.
20. Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. Biometrics 1994;50:350–357.
21. Miettinen OS. Theoretical Epidemiology: Principles of Occurrence Research in Medicine. New York: John Wiley and Sons, 1985.
22. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. Am J Epidemiol 1992;135: 1019–1028.
23. Little RJ, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley and Sons, 1987.
24. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to use of ad hoc methods of correcting for missing values of confounding variables. Biometrics 1991;47:63–76.
25. Hsieh DA, Manski CF, McFadden D. Estimation of response probabilities from augmented retrospective observations. J Am Stat Assoc 1985;80:651–662.
26. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol 1980;112:467–470.
27. Miettinen OS. Estimability and estimation in case-referent studies. Am J Epidemiol 1976;103:226–235.
28. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol 1986;123:174–184.
29. Stijnen T, van Houwelingen HC. Relative risk, risk difference and rate difference models for sparse stratified data. Stat Med 1993;12:2285–2293.
30. McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed. London: Chapman and Hall, 1989.
31. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. Am J Epidemiol 1992;135:1029–1041.
32. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 1977;39:1–38.
33. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical Recipes: the Art of Scientific Computing. Cambridge: Cambridge University Press, 1986.
34. Hartge P, Cahill JJ, West D, Hauck M, Austin D, Silverman D, Hoover R. Design and methods in a multi-center case-control interview study. Am J Public Health 1985;74:52–56.
35. Silverman DT, Levin LI, Hoover RN, Hartge P. Occupational risk factors of bladder cancer in the United States. I. White men. Am J Epidemiol 1986;123:174–184.
36. Weinberg CR. Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. Am J Epidemiol 1986;123:162–173.
37. Wacholder S, Carroll RJ, Pee D, Gail MH. The partial questionnaire design for case-control studies. Stat Med 1994;13:623–634.
38. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 1982;115:119–128.
39. Weinberg CR, Wacholder S. The design and analysis of case-control studies with biased sampling. Biometrics 1990;46:963–975.
40. Flanders WD, Greenland S. Analytic methods for two stage case-control studies and other stratified designs. Stat Med 1991;10:739–747.
41. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. Am J Epidemiol 1992; 135:1042–1050.
42. Wacholder S, Gail MH, Pee D. Selecting an efficient design for assessing exposure disease relationships in an assembled cohort. Biometrics 1991;47: 63–76.